MULTIFUNCIONALIDAD DE LOS CORPUS PARALELOS, EJEMPLIFICADA CON EL CORPUS ALEMÁN / ESPAÑOL PaGeS

Multifunctionality of parallel corpora, exemplified by German-Spanish corpus PaGeS

Irene Doval, Tomás Jiménez
Universidade de Santiago de Compostela

Resumen

En este capítulo se muestra cómo un corpus paralelo, ejemplificado con el corpus bilingüe alemán / español PaGeS, puede convertirse en una herramienta multifuncional apta para muy diferentes tipos de usuarios, si su composición, recuperación de información, posibilidades de búsqueda y visualización cumplen ciertos criterios. El corpus PaGeS consta de dos partes, un núcleo y unos suplementos. En este trabajo se describen los diferentes pasos seguidos en la construcción del corpus nuclear, una colección de narrativa contemporánea española y alemana. Esta descripción incluye la preparación manual de los textos, el alineado automático y la revisión manual. Se explica el acceso y visualización de los datos, así como los diferentes niveles de búsqueda adaptados a distintos usuarios. Finalmente se hace un balance y se esbozan las líneas de desarrollo futuro.

Palabras clave: lingüística de corpus, corpus paralelo, alineado del corpus, visualización del corpus

Abstract

This chapter shows how a parallel corpus, exemplified by the German-Spanish bilingual corpus PaGeS, can become a multifunctional tool, suitable for many different types of users, provided that its composition, information retrieval and the possibilities of searching and visualization meet certain criteria. The PaGeS corpus consists of two parts,

the core corpus and the supplements. In the present study we describe the different steps taken in the construction of the core corpus, which consists of a collection of contemporary Spanish and German narrative texts and their translations. This description includes the manual preparation process of the texts and the manual and automatic procedure of sentence alignment. It explains the access and the visualization of the data, as well as different search levels according to the needs of different users. The study ends with an evaluation and outline of the next steps to be taken in the future.

Keywords: corpus linguistics, parallel corpora, corpus alignment, corpus visualization

1. Introducción: Los corpus paralelos

Desde los años 90 las investigaciones lingüísticas basadas en corpus se han generalizado como el método estándar, revolucionando todas las disciplinas lingüísticas. Los primeros corpus se remontan ya a la década de 1960, eran exclusivamente monolingües y mayoritariamente de textos ingleses. Durante más de dos décadas, los corpus continuaron siendo monolingües, hasta que la creación del Corpus Paralelo Inglés/Noruego[1] (ENPC) (Johansson & Hofland 1994) y su proyecto hermano Corpus Paralelo Inglés/Sueco (Aijmer & Altenberg 1996: 79 ss.), a comienzos de los años 90, marcaron el comienzo de la era de los corpus paralelos y fueron determinantes en su ulterior evolución. Siguiendo este modelo se compilaron varios corpus en los años siguientes (Hasselgard 2015: 4). Este rápido desarrollo de los corpus paralelos le permitió a Lars Borin, ya en 2002 afirmar que «in the last decade or so, parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics, itself a fairly young discipline» (Borin 2002: 1). Desde entonces se sucedieron conferencias y workshops dedicados exclusivamente a los corpus paralelos y comparables[2], a su creación, compilación, anotación y procesado. Asimismo, una amplia variedad de corpus paralelos fueron creados para diferentes lenguas y con diferentes objetivos.

Este es el caso de los recursos derivados de textos producidos por las diferentes instituciones de la Unión Europea (*vid.* Steinberger *et al.* 2014 para una visión general). Estos están incluidos, por ejemplo, en el corpus Multilin-

[1] Hubo corpus paralelos anteriores, como el de Filipovic o el *Canadian Hansard Corpus*, pero permanecieron como iniciativas aisladas sin continuidad. Para una visión más detallada sobre la evolución de los corpus paralelos, *vid.* Doval & Sánchez (2019: 1-19).

[2] Según la terminología comúnmente aceptada (McEnery & Hardie 2012: 20), en los corpus multilingües se distingue entre corpus comparables y paralelos. A diferencia de los corpus paralelos, los textos de un corpus comparable no son traducciones unos de otros, sino que son textos monolingües en diferentes idiomas que comparten tema, tipología textual y registro con un origen y un alcance similares.

gwis (Clematide *et al.* 2016), una herramienta de búsqueda desarrollada en la Universidad de Zurich que cubre los debates del Parlamento Europeo en siete lenguas, incluyendo alemán y español. Forman parte también de la gran colección de corpus paralelos integrados en el proyecto Opus de Jörg Tiedemann (Tiedemann 2012). Además de documentos administrativos de las instituciones europeas, Opus contiene también textos periodísticos y algunas colecciones menores de diferentes fuentes en línea, como subtítulos y documentación técnica. Otro corpus paralelo multilingüe es InterCorp, compilado en la Universidad Carolina de Praga (Čermák 2019). Abarca 40 lenguas, entre ellas el alemán y español. Además de los textos de la UE y subtítulos, incluye también textos de ficción, utilizando el checo como la lengua pivote para el alineado.

Por último, hay que mencionar un recurso que tiene un amplísimo uso, Linguee, que es una herramienta en línea que combina ciertas prestaciones de diccionarios bilingües con ejemplos de uso alineados paralelamente, acercándose también a lo que es una memoria de traducción. Actualmente cubre 25 lenguas, según informaciones propias y, aunque dispone de una cierta variedad de textos, la mayoría están vinculados al tipo de texto administrativo o comercial de la Unión Europea.

El objetivo de este capítulo es presentar el corpus paralelo alemán / español, PaGeS[3], mediante una descripción de su contenido y de sus características distintivas (§ 2); la sección 3 está destinada a explicar el procesado, segmentación y alineado de los textos, mientras que la indexación, posibilidades de búsqueda y visualización de los datos están comentados en la sección 4. Por último, en el § 5 se hace un balance en el que se incluyen las funcionalidades previstas no implementadas, así como potenciales ampliaciones futuras.

2. El corpus paralelo alemán/español PaGeS: composición

Como se ha indicado en el §1, la mayoría de los recursos multilingües disponibles para el par de lenguas alemán/español se limitan a variedades textuales específicas, principalmente lengua jurídica, administrativa o técnica. Además, en la inmensa mayoría de los casos no se puede determinar con precisión ni la lengua original, ni el proceso de traducción llevado a cabo. Así que, tanto por el tipo de textos como por las características de su producción, estos recursos, aunque indudablemente valiosos, adolecen de limitaciones como

[3] Este corpus que se encuentra disponible en línea desde 2016 (www.corpuspages.eu) está siendo elaborado por el equipo de investigación SpatiAlEs, de la Universidad de Santiago de Compostela, dirigido por Irene Doval. Este grupo consta de miembros de las universidades de Santiago de Compostela, Salamanca, Complutense de Madrid y Valladolid. El proyecto de elaboración del corpus está siendo financiado por el Ministerio de Economía y Competitividad (FFI2013-42571-P y FFI2017-85938-R).

herramientas multifuncionales; en efecto, los corpus precedentes presentan ciertas desventajas tanto para la investigación en lingüística contrastiva y traducción como para la enseñanza y aprendizaje de lenguas extranjeras, tareas fundamentales de nuestro grupo de investigación (Doval 2017b: 128).

La lengua comercial, administrativa y jurídica presenta poca variedad léxica y morfosintáctica en las estructuras gramaticales más empleadas, ya que se tiende al uso de fórmulas estereotipadas. Asimismo, este tipo de discurso no resulta adecuado para la enseñanza de la lengua con fines no específicos, por presentar un nivel de dificultad demasiado elevado.

Para la investigación en lingüística contrastiva y para el estudio de fenómenos relacionados con la lengua traducida, es indispensable identificar inequívocamente la lengua original y la lengua traducida, esto es, determinar cuál es la lengua fuente y cuál la lengua meta. Además, resulta también fundamental conocer el proceso de traducción, esto es, si se trata de una traducción directa entre el alemán y el español o, si por el contrario, se trata de una traducción indirecta a través de una tercera lengua.

Por último, hay que señalar que para cualquier investigación lingüística o recurso didáctico es condición indispensable la calidad de los materiales que constituyen la base empírica del trabajo. Los recursos mencionados, a excepción de los textos de la Unión Europea, no proporcionan estándares de calidad ni para los textos originales ni para las traducciones, ya que no han sido sometidos a controles de calidad comprobables.

Estas limitaciones de los recursos existentes han constituido la motivación inmediata para la creación del corpus PaGeS. Dado que las actividades del grupo de investigación se centran en la lingüística contrastiva y en la enseñanza de lenguas, resulta imprescindible asegurar la calidad del material, tanto con respecto a los textos originales como a su traducción. Para garantizarla, los textos tienen que haber pasado necesariamente algún control de calidad. Por otro lado, hay que tener en cuenta que la compilación de un corpus paralelo, frente a uno monolingüe, tiene enormes limitaciones en cuanto al material disponible, ya que la inmensa mayoría de los textos no se traducen y, cuando lo están, solo una pequeña parte pasa algún tipo de control. Por ello, la única vía disponible es acudir a materiales publicados por editoriales reconocidas, donde originales y traducciones hayan sido sometidos a un estricto control de calidad (Doval 2017b, 2018).

Por estos motivos, en PaGeS se ha compilado un corpus nuclear de textos narrativos escritos después de 1960, aunque con un claro predominio de obras a partir del año 2000. En cuanto al género, están integrados tanto textos de ficción como una pequeña parte de no ficción, ya que constituyen

la mayor parte de los recursos bilingües disponibles. Actualmente (abril de 2019) están incluidos textos de 140 obras originales (en español o alemán, además de una pequeña parte en una tercera lengua, *vid.* figura 1) y sus traducciones, con un tamaño total de más de 25 millones de palabras, unos 28 millones de tokens y 891.265 bisegmentos, es decir, pares de unidades alineadas (oraciones o segmentos más pequeños). En las obras literarias se ha tratado de atender a la variedad de géneros, con cierto predominio de la literatura infantil, así como a la variedad dialectal. En los textos españoles hay una amplia muestra de autores americanos y, en los textos alemanes, los autores suizos y austríacos están también representados. Entre las obras de no ficción se encuentran ensayos políticos e históricos, así como prosa de divulgación científica. Por otro lado, se ha velado por el equilibrio en las direcciones de la traducción, por lo que el corpus ofrece una composición proporcionada en cuanto a la lengua original, tal como se recoge en la figura 1. La tabla 1 muestra la composición del corpus nuclear, disponible ya en línea, con la indicación del número de obras y de palabras distribuidas según la lengua original.

Lengua	Obras	Palabras	Bisegmentos
Alemán Original	66	5 354 413	406 029
Español Traducción < Alemán	66	5 601 475	406 029
Español Original	56	5 221 244	332 955
Alemán Traducción < Español	56	5 145 174	332 955
Alemán Traducción < 3ª Lengua	18	2 101 211	152 347
Español Traducción < 3ª Lengua	18	2 157 039	152 347
Total	140 (x 2)	25 580 556	891 265 (x2)

TABLA 1. Composición del corpus nuclear de PaGeS (abril de 2019)

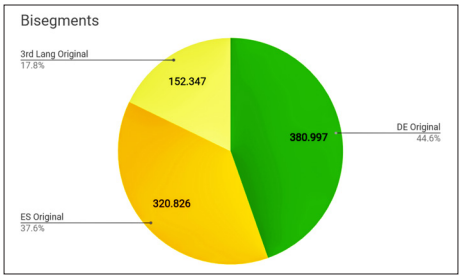

FIGURA 1. Distribución de los bisegmentos por lengua original (abril 2019)

Además de este material nuclear, el corpus PaGeS ofrece otros materiales complementarios, que no reúnen todos los requisitos enumerados anteriormente, pero que, para determinados fines, pueden suponer un complemento valioso. En el momento actual estos materiales están compuestos por las versiones alemana y española del corpus Europarl y de los discursos TED.

El Europarl (Release v7) está constituido por las actas del Parlamento Europeo de los años 1996 a 2011. El corpus ha sido extraído y alineado automáticamente en corpus paralelos con la versión inglesa por Koehn (Koehn 2005). La versión utilizada en PaGeS para el par alemán/español procede de Opus (Tiedemann 2012). La lengua original no está consignada en la mayoría de las ocasiones y suponemos que, conforme a la praxis habitual de las instituciones europeas, se ha partido de la versión inglesa para la traducción a las respectivas lenguas, por lo que una traducción directa entre español y alemán constituye un hecho excepcional.

Los textos de TED proceden de las traducciones al alemán y al español de las transcripciones de las charlas TED, todas ellas originalmente pronunciadas en inglés. Las traducciones han sido llevadas a cabo por traductores voluntarios y sometidas a un proceso de revisión. Usamos la versión en xml ofrecida por Web Inventory of Transcribed and Translated Talks (WIT3) (Cettolo /Girardi /Federico 2012). Posteriormente han sido alineadas, usando el LF-Aligner y sometidas a diversos procesos de limpieza automática antes de su indexación. La tabla 2 presenta las estadísticas correspondientes a estos suplementos.

	Europarl		**TED**	
	Palabras	**Bisegmentos**	**Palabras**	**Bisegmentos**
Alemán	44 303 154	1 882 959	4 061 184	234 328
Español	49 860 899		4 248 416	

Tabla 2. Europarl y TED integrado en PaGeS

3. Procesado, segmentación y alineado de los textos

Esta sección se refiere exclusivamente al corpus nuclear, esto es, la parte del corpus PaGeS que procede de obras narrativas publicadas en editoriales. Después de la selección de textos, estos han de ser preparados para el alineado. Para ello ambas versiones se almacenan como archivos txt con la codificación común UTF-8. En una primera fase se trata de reducir lo máximo posible el ruido y de revisar los textos para que el texto fuente original y el meta sean lo

más paralelos posibles, a fin de conseguir unos resultados más precisos. De ahí que se eliminen todos los pasajes que no estén presentes en ambas versiones o que no pertenezcan al cuerpo de la obra: información bibliográfica (que se almacena en un archivo aparte como metadatos), dedicatorias, epígrafes, prefacios, apéndices, notas o bibliografías.

A continuación, los textos son revisados cuidadosamente a fin de detectar eventuales errores provocados por el proceso de digitalización de las obras más antiguas, pues las más recientes ya han sido creadas digitalmente. Se marca la estructura interna de la obra, esto es, la eventual división en partes o capítulos. Por último, se recogen y almacenan en un archivo aparte los metadatos descriptivos de cada obra (*cfr*. Doval 2018: 183).

El alineado es, obviamente, una fase crucial en la construcción de un corpus paralelo, y es definido por Tiedemann (2011: 123) como «a process of making symmetric correspondences explicit in order to enable further processing of parallel resources». Este conjunto de correspondencias entre el texto fuente y el meta es lo que forma el llamado bitexto. Las unidades de alineado del bitexto dependen del nivel de detalle de la segmentación que se considere: parágrafos, oraciones o palabras. Actualmente en los recursos paralelos el alineado a nivel de oración constituye un estándar y es el más común en los corpus paralelos (Tiedemann 2011: 37). Por ello, en PaGeS nos hemos centrado en la oración como la unidad básica de alineado.

En este proceso se combinan dos tareas, una previa de segmentación de los textos en oraciones y la posterior vinculación de esos segmentos con sus correspondencias en la otra lengua para formar bisegmentos. La segmentación se lleva a cabo para cada lengua independientemente y el algoritmo realiza en principio un corte por cada puntuación fuerte, esto es, punto, signo de interrogación o de admiración final[4] (Zanettin 2012: 158). Ahora bien, como la puntuación en alemán y español no siempre es coincidente, esta segmentación inicial es posteriormente rectificada, caso necesario, en el proceso de alineado (*vid. infra*). Además, el signo de puntuación punto (.) es ambiguo, ya que no siempre indica el final de una oración, sino que se utiliza como marcador de abreviaturas (en alemán y español) y en alemán también como marcador de número ordinal, tal como ilustra el siguiente ejemplo:

(1) Dr. Kaltensee feiert heute in U.S.A. seinen 30. Geburtstag. [El Dr. Kaltensee celebra hoy en USA su 30º cumpleaños]

[4] Más precisamente (Scott 2010, citado en Zanettin 2012: 158) puntualiza: «A sentence ends if a full-stop, question-mark or exclamation-mark (.?!) is immediately followed by one or more word separators and if the next non-punctuation symbol is a capital letter A..Z or an accented capital letter, a number or a currency symbol».

Se hace, por tanto, necesario desambiguar el punto de final de oración del que forma parte de la palabra, como en las abreviaturas o en los números ordinales. Para ello, se integra una lista finita de abreviaturas comunes para alemán y español en el segmentador. Evidentemente, esta lista no es exhaustiva, ya que continuamente se crean abreviaturas nuevas u ocasionales.

Tiedemann (2011: 9) subraya la importancia de la segmentación para la precisión del alineado posterior: «The importance of segmentation is often ignored in the literature on text alignment. However, it plays a crucial role in the success of the algorithm». Efectivamente, buen número de los fallos en el alineado se deben a fallos en la segmentación previa. Esto ocurre especialmente en los casos de puntuación no coincidente, tal como ilustra el ejemplo siguiente, en el que en español hay dos puntos y no se han segmentado, mientras que en alemán hay un punto y se ha introducido, por tanto, una segmentación (Pérez Reverte 2012, cap. 3).

	Es kostete nichts, freundlich zu sein.
No costaba nada ser amable: invertir de cara al futuro.	Es war eine Investition in die Zukunft.

Tabla 3. Ejemplo de puntuación divergente

Tras realizar varios tests con diferentes herramientas de alineado[5], hemos optado por LF-Aligner, ya que alcanzó en ellos una mayor precisión. Está basado en el algoritmo de Hun-Align[6] (Varga 2012: 92-119), un enfoque híbrido que combina el algoritmo basado en la semejanza de longitud de las oraciones con el de las correspondencias léxicas (Tóth *et al.* 2008). Como las correspondencias léxicas se derivan de forma automática, el LF-Aligner no requiere de un léxico externo.

El alineado de estos segmentos resultantes se ejecuta en cuatro fases: 1) el archivo input es un texto «tokenizado» y segmentado en oraciones en las dos lenguas; 2) este texto se alinea usando una versión modificada del modelo de Brown *et al.* (1993), basado en la longitud de las respectivas oraciones; 3) se construye un diccionario automático bilingüe basado en este primer alineado y 4) finalmente, se realinea el texto en un segundo paso, tomando en consideración la semejanza léxica, proporcionada por el diccionario automático (Toth *et al.* 2008: 470).

[5] Se han realizado tests con los siguientes alineadores; ABBY Aligner, bitext2tmx y Vanilla aligner (Danielsson & Ridings 1997).

[6] Hun-Align es una elección común entre los creadores de corpus paralelos multilingües. Por ejemplo se usó para el alineado de Intercorp (Čermák 2019), en la plataforma Multilingwis (Clematide *et al.* 2016), o en el alineado de JRC-Acquis (Steinberg *et al.* 2014: 1) entre otros.

La precisión del alineado depende, en primer lugar, del tipo de textos; concretamente, los textos ficcionales, que forman el núcleo del corpus PaGeS, presentan mayores problemas que otro tipo de textos, como los administrativos o técnicos, tal como señala Zanettin (2012: 155):

> Some text types are better suited to automatic alignment than others. Typically, parallel collections of technical, legal or otherwise official documents such as transcriptions of parliamentary proceedings, especially in related languages, are, in this respect, much better material than journalistic and fictional texts. While the former usually present the same formal structure and contain the same number of paragraphs and sentences as well as fixed and standard punctuation, the latter often score low on these features.

Además, dentro de los textos literarios, el grado de correspondencia varía dependiendo del autor, del traductor, de los propios textos y de la dirección de la traducción.

Obviamente, una correspondencia 1:1 entre la lengua fuente y la meta no es siempre posible, puesto que en el proceso traductológico las oraciones pueden ser divididas (1:2), fusionadas (2:1) o reordenadas; además, el traductor puede optar por omitir o insertar oraciones o pasajes de texto (Tiedemann 2011: 9; Varga 2012: 94). Las tablas 4 y 5 muestran algunos de estos casos. Además, tal como se ha señalado anteriormente, PaGeS incluye también bitextos en los que ambas versiones alemana y española son traducciones de una tercera lengua y estos textos son particularmente problemáticos, ya que han sufrido dos procesos de traducción independientes (Doval 2018: 186).

	Im hinteren Teil sorgt eine Reihe von Leuchten für gleichmäßiges, gedämpftes Licht.
Al fondo, frente a ocho filas de asientos ocupados por el público, una instalación de luz artificial amortiguada, uniforme, ilumina la mesa de juego situada en una tarima y un gran tablero mural de madera que hay en la pared, junto a la mesa del árbitro, donde un ayudante de éste reproduce el desarrollo de la partida.	Auf einem Podest steht dort der Spieltisch vor acht mit Zuschauern besetzten Stuhlreihen, und an der Wand hängt ein großes Schachbrett, auf dem der Schiedsrichterassistent die Spielzüge nachstellt.

Tabla 4. Ejemplo de mal alineado debido a una correspondencia 1:2

	Sie lachte laut auf.
Ella soltó una carcajada viva y fuerte.	Ein fröhliches und herzhaftes Lachen.
Una risa sana. —Exacto —asentía, siguiéndole la corriente con buen humor—.	„Genau«, bestätigte sie. „Wie haben Sie das erraten?«

Tabla 5. Ejemplo de mal alineado debido a reordenación

Después del alineado automático, como se ha dicho, se valida manualmente el resultado. Solo de esta manera se puede conseguir una tasa de error inferior al 0,5 %. Para ello se procede en tres fases. Primero se seleccionan los segmentos de más de 350 caracteres, ya que han de ser divididos para poder ser procesados. Para ello se insertan manualmente marcas (breaks)
 en lugares adecuados del texto de ambos segmentos, que luego son divididos automáticamente. En un segundo paso se localizan los alineados vacíos, es decir, los segmentos no emparejados. En este caso, puede tratarse de un alineado erróneo o de eliminaciones o inserciones en el texto traducido. Si el segmento está desalineado, se hacen las correcciones necesarias. Si el segmento no ha sido traducido, se inserta en la celda vacía la marca [n_t_s] (=non translated segment). Si se ha añadido el segmento en el texto traducido, se inserta la marca [a_s_t] (=texto añadido en la traducción). Finalmente, para minimizar el trabajo manual, nos centramos en los bisegmentos en los que, debido a desfases de longitud entre el segmento fuente y el meta, es más probable que haya errores. Para identificarlos calculamos el cociente de la suma de caracteres del bisegmento y la diferencia de caracteres entre el segmento fuente y meta. Luego aplicamos esta ratio para ordenar los bisegmentos. Los errores tienden a ocurrir en los bisegmentos donde el rango de valores de la ratio está entre -5 - 5. De esta manera la comprobación manual de los resultados del alineado se realiza de forma más eficiente y requiere menos tiempo. Este procedimiento es un compromiso entre lo que sería deseable y lo que es factible, asegurando asimismo un alto nivel de precisión.

4. Búsqueda y visualización de los resultados

Una vez que los archivos alineados han sido revisados, se indexan y se gestionan a través del motor de búsqueda general Apache-Solr (Versión 7.5.0), una potente y muy rápida plataforma de búsqueda de código abierto escrita en Java, que cubre una amplia gama de funcionalidades.

Como se mencionó antes, PaGeS pretende ser una herramienta realmente polivalente, útil para grupos de usuarios muy diversos, desde investigadores en lingüística y traducción hasta lexicógrafos, expertos en PLN, pasando por usuarios no especialistas, sean ocasionales o habituales, así como estudiantes de alemán o español. Para ello es esencial proporcionarles una interfaz adecuada para la visualización y recuperación de los datos, esto es, los textos del corpus, los metadatos y las eventuales anotaciones lingüísticas. Para lograr este objetivo, la interfaz debe ofrecer una serie de funcionalidades básicas (Doval *et al.* 2019: 115). Por un lado, la búsqueda ha de ser:

(a) Rápida, ya que se prevé un aumento significativo del tamaño del corpus PaGeS. El motor de búsqueda debe permitir realizar búsquedas de forma rápida y eficaz a través de grandes cantidades de datos lingüísticos.
(b) Amigable. El lenguaje de la consulta debe ser lo más sencillo posible. Un lenguaje de consulta avanzado y más complejo solo se muestra si es necesario. Además, deben aprovecharse los hábitos de búsqueda de los usuarios de Internet y, por lo tanto, el lenguaje de consulta de Google debe servir de modelo.
(c) A varios niveles: El sistema de búsqueda debe permitir consultas a través de múltiples capas de anotación lingüística, como la lematización, el etiquetado de clases de palabras o la anotación sintáctica.

Por otro lado, los resultados de las consultas han de presentarse en un formato fácil de leer. Los segmentos fuente y meta deben mostrarse uno al lado del otro, y tanto la palabra o frase de búsqueda como su equivalente potencial deben ser resaltados.

Por ello y para satisfacer las necesidades de los usuarios mencionados anteriormente, hemos diseñado una búsqueda en tres niveles. La primera, cuya interfaz se muestra en la figura 3, es la búsqueda simple o estándar. En este caso, el usuario solo tiene que introducir en el campo de búsqueda el término de la búsqueda (una palabra o una frase) en alemán o español. En este tipo de consultas, la lematización se aplica por defecto. Con las consultas multipalabra, se encuentran todas las palabras de búsqueda dentro de una distancia específica. Al igual que en la búsqueda de Google, si el término se introduce entrecomillado (por ejemplo «pasar de largo») la búsqueda solo devuelve resultados que coincidan exactamente con la forma de la palabra o frase introducida (*vid*. figura 4). Por defecto, las búsquedas solo incluyen el corpus nuclear, aunque el usuario puede seleccionar la casilla correspondiente a los textos complementarios (Europarl o TED).

La forma más popular de mostrar los resultados de una búsqueda en un corpus es en forma de concordancia y el formato de concordancia más común es la concordancia KWIC (Key Word in Context) (Wynne 2008: 706), es decir, la palabra objeto de la consulta está en una posición central con todas las líneas alineadas verticalmente alrededor de ella. Esta presentación de los resultados, debidamente ordenada, es muy útil en corpus monolingües para visualizar patrones de uso. Sin embargo, en los corpus bilingües, el formato KWIC no puede considerarse amigable, ya que una de las principales aplicaciones de estos corpus es encontrar rápidamente posibles equivalentes

de un término de búsqueda. La figura 2 ilustra este tipo de presentación en un corpus alineado por AntPConc.

FIGURA 2. Visualización kwic de AntPConc

Por esta razón, nos decidimos por la visualización de los resultados de la consulta en una tabla de dos columnas, donde una columna corresponde a los textos fuente y la otra a los textos meta. El término de búsqueda se muestra en una celda con algún contexto y se resalta en negrita. Dependiendo de si el término de la búsqueda se encuentra en el texto original o en la traducción, se muestra en una u otra columna. Los resultados en los textos originales se muestran primero en la columna izquierda, los de las traducciones en la columna derecha. El segmento correspondiente se visualiza en la misma línea, pero en una celda de la otra columna. En cada consulta, se informa sobre el número de ocurrencias y el número de páginas, así como el número de la página actual. Las figuras 3 y 4 proporcionan ejemplos del menú de búsqueda estándar y algunas de sus características:

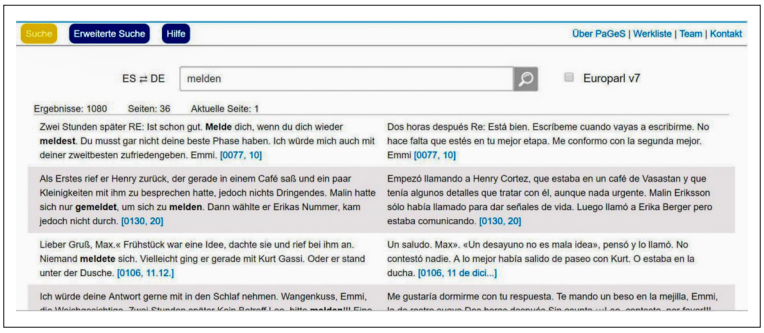

FIGURA 3. Búsqueda simple en el corpus PaGeS

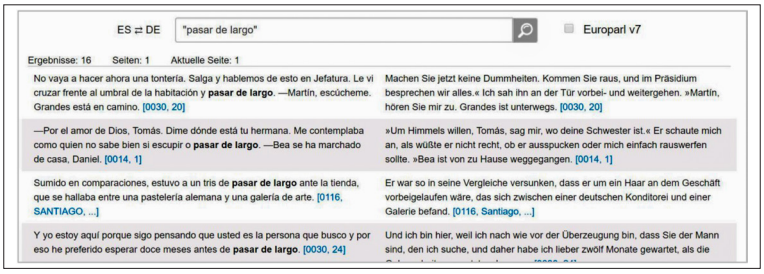

FIGURA 4. Búsqueda multipalabra exacta en el corpus PaGeS

Al final de la tabla hay un link que permite navegar a través de las páginas y descargar los resultados de la búsqueda en formato CSV a los usuarios registrados En cada pasaje se ofrece, entre corchetes, información relativa al ID de la obra, así como la parte y el capítulo en el que se encuentra. Haciendo clic sobre el ID de la obra, el usuario puede ver un contexto lingüístico mayor, seleccionando el número de segmentos antes y después de la ocurrencia. Además, en esta pantalla se muestra la información bibliográfica completa de la obra, tal como muestra la figura 5.

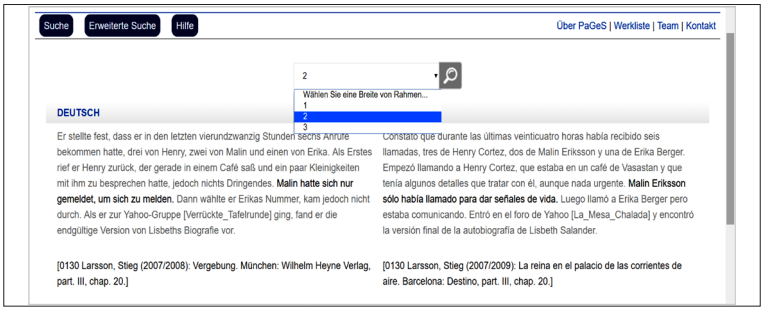

FIGURA 5. Ampliación de contexto e información bibliográfica

El segundo nivel de búsqueda es el avanzado. En este nivel, el usuario puede restringir el alcance de su búsqueda aplicando filtros operativos mediante menús desplegables. Puede limitar la búsqueda por autores, obras, años de publicación o género. También puede seleccionar si la búsqueda ha de operar solo en originales o en traducciones. Además, puede buscar un término solo cuando se traduce de una manera concreta. La figura 6 muestra un ejemplo de búsqueda avanzada.

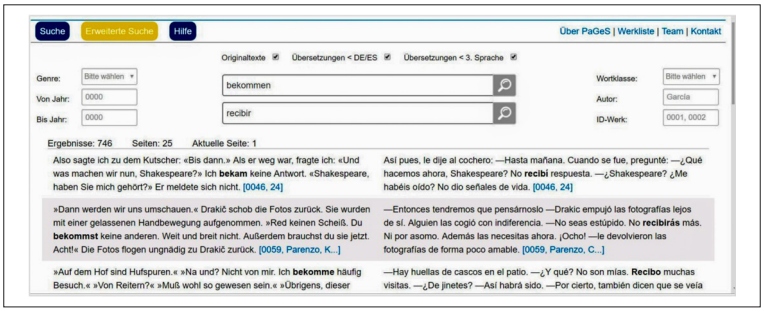

FIGURA 6. Búsqueda avanzada en PaGeS

El último nivel de búsqueda es el más complejo, y actualmente está parcialmente disponible a través de la interfaz de búsqueda estándar. Este nivel proporciona al usuario el comando completo de la sintaxis de consulta utilizada en la herramienta de indexación subyacente Solr. Esta soporta búsquedas usando expresiones regulares. El término de búsqueda debe ir precedido de [SS] (Solr Search). La expresión de búsqueda ha de ser construida palabra a palabra, y para cada palabra, es posible especificar varios parámetros. Este nivel permitirá, además, la búsqueda combinada de palabras o frases y etiquetas de clases de palabras u otras anotaciones, cuando estén implementadas. Esta búsqueda formal permite la ejecución de consultas muy precisas, pero no es particularmente fácil de usar. Está dirigida a usuarios más exigentes, como los investigadores en lingüística contrastiva o traducción, que suelen necesitar un subconjunto muy específico de resultados, solo posible con consultas complejas que incluyan un gran número de parámetros.

5. BALANCE Y PERSPECTIVAS

En los apartados anteriores hemos presentado las características distintivas del corpus PaGeS y los procesos implicados en su compilación e indexación. Los tres niveles de búsqueda están adaptados a las necesidades de los distintos grupos de usuarios, lo que hace del corpus PaGeS un recurso multifuncional con un enorme potencial. Sus aplicaciones concretas en la lingüística contrastiva y el aprendizaje de idiomas se están explotando actualmente en nuestro grupo de investigación (Doval 2018, Lübke & Liste Lamas 2018).

En todo caso tenemos previsto introducir nuevas funcionalidades. Las más inmediatas y que ya están en parte realizadas son el alineado de palabras y el etiquetado de clases de palabras. El alineado de palabras es muy útil, especialmente para los estudiantes, ya que facilita la identificación a primera

vista de la palabra equivalente del término de búsqueda en la otra lengua (Volk *et al.* 2014). Actualmente se están realizando pruebas con los siguientes alineadores: Giza++, Berkeley Aligner y eflomal.

En cuanto al etiquetado con clases de palabras ya ha sido realizado para el corpus nuclear, utilizando para el español el etiquetador Freeling y para el alemán el TreeTagger (*vid.* Doval 2017a). La inclusión de esta información en el indexado permitirá realizar consultas complejas, en las que se combine la búsqueda de la cadena y la categoría.

Por otro lado, ya ha sido iniciada la construcción de otros dos corpus paralelos: español/holandés, el corpus PaDeS, y español/chino, el corpus PaZheS, con los mismos criterios de diseño y características de consulta. No están todavía en línea, pues no disponen en el momento actual de la suficiente cantidad de material paralelizado mínimo para que pueda resultar útil. Está prevista su puesta en línea en cuanto se alcancen los cinco millones de palabras en el corpus nuclear.

Por último, somos muy conscientes de algunas deficiencias actuales de nuestro motor de búsqueda, como la limitación en la ordenación de los resultados o la selección de colocaciones. De todos modos y, a pesar de la existencia de otros corpus paralelos, PaGeS presenta una serie de características distintivas, entre las que cabe destacar: el tipo de textos utilizado de gran variedad léxica y gramatical, la alta calidad de originales y traducciones, el equilibrio en la bidireccionalidad, la revisión manual de los procesos automáticos y, finalmente, su disponibilidad en línea y su facilidad de uso. Todas estas características lo convierten en un recurso multifuncional adecuado para múltiples aplicaciones, desde la investigación contrastiva, pasando por los estudios de traducción hasta el aprendizaje y enseñanza de lenguas extranjeras.

Recursos electrónicos

ABBY Aligner: https://www.abbyy.com/en-eu/aligner/

AntPConc: https://www.laurenceanthony.net/software/antpconc/

Apache-Solr (Versión 7.5.0) http://lucene.apache.org/solr/

Berkeley Aligner: https://github.com/mhajiloo/berkeleyaligner

Bitext2tmx http://bitext2tmx.sourceforge.net/

Eflomal: Efficient Low-Memory Aligner https://github.com/robertostling/eflomal

Freeling: http://nlp.lsi.upc.edu/freeling/node/1

Giza++: https://github.com/moses-smt/giza-pp

LF-Aligner: http://sourceforge.net/projects/aligner/

Linguee: Linguee GmbH (2010): Linguee.com: The web as a dictionary. http://www.linguee.com

Multilingwis: https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo/

Opus: http://opus.nlpl.eu/

TED: www.ted.com

TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

WIT3: Web Inventory of Transcribed and Translated Talks https://wit3.fbk.eu/

Referencias bibliográficas

Aijmer, Karen, Bengt Altenberg & Mats Johansson (eds.) (1996): *Languages in contrast: papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*. Lund: Lund University Press, pp. 73-85.

Borin, Lars (2002): «...and never the twain shall meet?» in Lars Borin (ed.): *Parallel corpora, parallel worlds: selected papers from a symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam: Rodopi, pp. 1-43. https://doi.org/10.1163/9789004334298_002

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer (1993): «The mathematics of statistical machine translation: parameter estimation», *Computational Linguistics* 19/2, pp. 263-311.

Čermák, Petr (2019): «InterCorp: Parallel corpus of 40 languages», in Irene Doval & M. Teresa Sánchez (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 93-101. https://doi.org/10.1075/scl.90.06cer

Cettolo, Mauro, Christian Girardi & Marcello Federico (2012): «WIT3: Web Inventory of Transcribed and Translated Talks», in *Proceedings of EAMT*, Trento, Italy, pp. 261-268.

Clematide, Simon, Johannes Graën & Martin Volk (2016): «Multilingwis: a multilingual search tool for multi-word units in multiparallel corpora», in Gloria Corpas Pastor (ed): *Computerised and corpusbased approaches to phraseology: monolingual and multilingual perspectives – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*. Geneva: Tradulex, pp. 447-455.

Doval, Irene & M. Teresa Sánchez Nieto (2019): «Parallel corpora in focus: an account of current achievements and challenges», in Irene Doval & M. Teresa Sánchez Nieto (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 1-19. https://doi.org/10.1075/scl.90.01dov

Doval, Irene, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas & Barbara Lübke (2019): «Corpus PaGeS: a multifunctional resource for language learning, translation and cross-linguistic research», in Irene Doval & M. Teresa Sánchez Nieto (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 103-121. https://doi.org/10.1075/scl.90.07dov

Doval, Irene (2017a): «POS-tagging a bilingual parallel corpus: methods and challenges», *Research in Corpus Linguistics* 5, pp. 35-46. https://doi.org/10.32714/ricl.05.03

Doval, Irene (2017b): «La construcción de un corpus paralelo bilingüe multifuncional», *Moenia: Revista lucense de lingüistica & literatura* 27, pp. 125-141.

Doval, Irene (2018): «Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache», *Revista de Filología Alemana* 26, pp. 181-197. https://doi.org/10.5209/RFAL.60148

Hasselgård, Hilde (2015): «Parallel corpora and contrastive studies», in *Proceedings of the international symposium on Using Corpora in Contrastive and Translation Studies, (UCCTS), 2010 Conference,* Edge Hill University, 27-29 July 2010. http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2010/Proceedings/papers/Hasselgard. pdf [consultado: 26 de junio de 2018].

Johansson, Stig & Knut Hofland (1994): «Towards an English-Norwegian parallel corpus», in Udo Fries, Gunnel Tottie & Peter Schneider (eds): *Creating and using English language corpora*. Amsterdam: Rodopi, pp. 25-37.

Koehn, Philipp (2005): «Europarl: a parallel corpus for statistical machine translation», in *Proceedings of Machine Translation Summit X*, Phuket, 13-15 September 2005. Vol. 5, pp. 79-86.

Lübke, Barbara & Elsa Liste Lamas (2018): *Raumrelationen im Deutschen: Kontrast, Erwerb und Übersetzung*. Tübingen: Stauffenburg.

McEnery, Tony & Andrew Hardie (2012): *Corpus linguistics: method, theory and practice*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Steinberger, Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco Benítez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro (2014): «An overview of the European Union's highly multilingual parallel corpora», *Language Resources and Evaluation* 48/4, pp. 679-707. https://doi.org/10.1007/s10579-014-9277-0

Tiedemann, Jörg (2011): *Bitext alignment*. [*s. l.*]: Morgan & Claypool Publishers. https://doi.org/10.2200/S00367ED1V01Y201106HLT014

Tiedemann, Jörg (2012): «Parallel data, tools and interfaces in OPUS», in *Proceedings of the 8th International Conference on Language Resources*

and Evaluation (LREC '2012), pp. 2214-2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [consultado: 20 de mayo de 2018].

Tóth, Krisztina, Richárd Farkas & András Kocsor (2008): «Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm», *Acta Cybern* 18, pp. 463-478.

Varga, Dániel (2012): *Natural language processing of large parallel corpora*. PhD thesis. Eötvös Loránd University, Budapest.

Volk, Martin, Johannes Graen & Elena Callegaro (2014): «Innovations in parallel corpus search tools», in *Proceedings of LREC, Reykjavik*. http://www.zora.uzh.ch/id/eprint/97282 /1/Volk_Graen_Callegaro_LREC_2014_v06.pdf

Wynne, Martin (2008): «Searching and concordancing», in Anke Lüdeling & Merja Kytö (eds): *Corpus linguistics: an international handbook*. Berlin: de Gruyter, pp. 706-737.

Xiao, Richard (2010): *Using corpora in contrastive and translation studies*. Cambridge Scholars Publishing.

Zanettin, Federico (2012): *Translation-driven corpora: corpus resources for descriptive and applied translation studies*. London: Routledge.